بسمه تعالی

تمرین سری چهارم درس یادگیری عمیق

پوریا محمدی نسب

۴۰۰۷۲۲۱۳۸

فهرست

سوال ۱.

مقاله زیر را با دقت مطالعه بفرمایید، ایده و روش پیشنهادی را به صورت کامل توضیح دهید.

Cubuk, Ekin D., Barret Zoph, Jonathon Shlens, and Quoc V. Le. "Randaugment: Practical automated data augmentation with a reduced search space." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 702-703. 2020.

داده افزایی (Data augmentation) روشی است که نشان داده است عملکرد مناسبی مخصوصا بر روی مدل های شبکه های عصبی عمیق دارد. اما تکنیک های داده افزایی معمولا با دو مانع اساسی روبه رو هستند. مانع اول افزایش پیچیدگی مجموعه آموزشی است که باعث بروز مشکل و افزایش هزینه محاسباتی برای فاز آموزش شبکه است. مانع و مشکل دوم با توجه به اینکه معمولا فضای جستجو شامل دو قسمت مجزا است تکنیک های داده افزایی توانایی افزایش قدرت تعمیم (generalization) را ندارند. روش پیشنهادی در این مقاله، دو مانع ذکر شده را تا حد امکان حل کرده است. تکنیک معرفی شده در این مقاله RandAugment نام دارد که از مزایای آن میتوان به کاهش بسیار زیاد فضای جستجو اشاره کرد که باعث میشود نیازی به تسک های نماینده مجزا برای آموزش مدل نداشته باشیم. بعلاوه، پارامتری بودن این روش کمک میکند تا قدرت تعمیم بتواند برای مدل ها و دیتاست های مختلف تنظیم شود.

در روش پیشنهادی، بر خلاف سایر روش ها که بر روی تعداد بیشتر تبدیلات (transformations) متمرکز بودند، تمرکز بر روی کاهش تعداد پارامترهای داده افزایی است به طوری که همچنان تا حد خوبی تنوع و وسعت در تصاویر وجود داشته باشد. به این منظور در الگوریتم پیشنهادی به جای استفاده از سیاست های آموزش دیده و احتمالات از یک رویه ی مستقل از پارامترها که همیشه یک تبدیل را به احتمال یکسانی انتخاب میکند استفاده میکنیم. برای مثال اگر N تبدیل مختلف داشته باشیم روش پیشنهادی K^n سیاست محتمل بیان میکند. سپس الگوریتم برای هر نوع تبدیل ضریبی در نظر میگیرد که این ضریب بین 0 تا 10 است و هر چه مقدار بزرگتری داشته باشد تاثیر و سهم بیشتری در داده افزایی خواهد داشت.

برای کاهش بیشتر فضای پارامترها، مشاهده شد که ضرایب تبدیل آموزش دیده از قبل رفتار مشابهی در حین آموزش مدل دارند و فرض شد که تنها یک اعوجاج M ممکن است برای پارامترسازی کل تبدیلات کافی باشد. در این مقاله ۴ روش برای مدیریت M در حین آموزش آزمایش شد. ۱) ثابت بودن ۲) رندم بودن ۳) افزایش به صورت خطی ۴) رندم بودن با افزایش حد بالا(سقف مقدار upper bound). الگوریتم پیشنهادی با دو پارامتر N (تعداد تبدیلات) و M(ضریب برای همه ی تبدیلات) کار میکند. افزایش مقدار هر دو متغیر منجر به خطی شدن مدل (regularization) میشود.

جدول زیر مقایسه نتایج سایر الگوریتم های داده افزایی و الگوریتم پیشنهادی که با نام RA مشخص است میباشد. تمام الگوریتم ها روی ۴ دیتاست ران با چند مدل شبکه عمیق مختلف مانند ResNet اجرا شده اند.

	baseline	PBA	Fast AA	AA	RA
CIFAR-10					
Wide-ResNet-28-2	94.9	-	-	**95.9**	95.8
Wide-ResNet-28-10	96.1	**97.4**	97.3	**97.4**	97.3
Shake-Shake	97.1	**98.0**	**98.0**	**98.0**	**98.0**
PyramidNet	97.3	**98.5**	98.3	**98.5**	**98.5**
CIFAR-100					
Wide-ResNet-28-2	75.4	-	-	**78.5**	78.3
Wide-ResNet-28-10	81.2	**83.3**	82.7	82.9	**83.3**
SVHN (core set)					
Wide-ResNet-28-2	96.7	-	-	98.0	**98.3**
Wide-ResNet-28-10	96.9	-	-	98.1	**98.3**
SVHN					
Wide-ResNet-28-2	98.2	-	-	**98.7**	**98.7**
Wide-ResNet-28-10	98.5	98.9	98.8	98.9	**99.0**

سوال ۲.

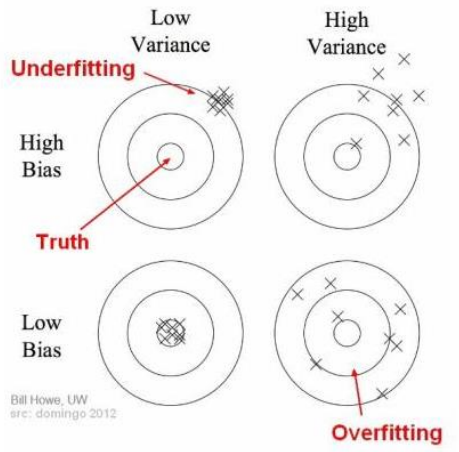
لطفا سوالات زیر را به صورت کامل پاسخ دهید

۲ – الف)

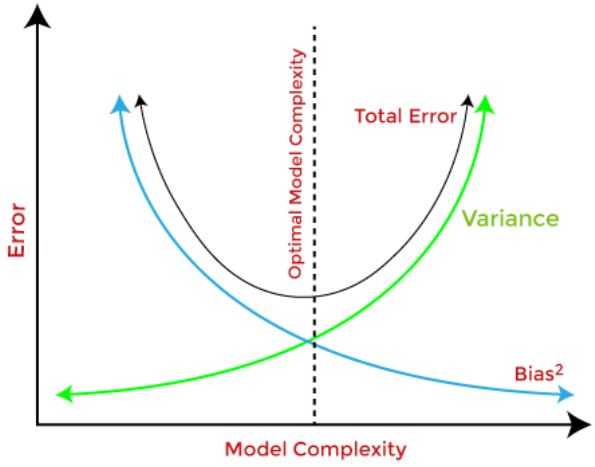
Bias-Variance tradeoff را با رسم شکل توضیح دهید.

توازن بین Bias و Variance به این معنی است که هر دو نوع خطا در حداقل مقدار خود باشند به طوری که مقدار خطای دیگر بالا نرود. برقراری این توازن بسیار مهم است زیرا نه تنها باعث میشود مدل آموزش دیده پیشبینی دقیقی داشته باشد بلکه از مشکلاتی مانند Overfitting و Underfitting جلوگیری میکند. به طور ساده Bias میانگین فاصله پیشبینی مدل و مقدار درستی است که سعی دارد آن را تخمین بزند. Variance میزان پراکندگی پیشبینی های انجام شده توسط مدل است.

این دو خطا چهار حالت مختلف میتوانند نسبت به هم داشته باشند که در شکل زیر قابل مشاهده است.



حال برای اینکه بتوانیم این توازن را برقرار کنیم، دو مقدار خطای Bias و Variance را با هم ادغام میکنیم و آن را total error مینامیم. بدیهی است هنگامی که total error مقدار کمینه داشته باشد، بهترین توازن بین دو خطا را خواهیم داشت. شکل زیر به صورت واضح تری این توازن را شرح میدهد.



۲- ب)

بایاس زیاد (high Bias) چطور قابل تشخیص است و برای مقابله با آن چه راه حل هایی وجود دارد؟

در قسمت الف این سوال دیدیم که بایاس زیاد باعث میشود مدل توجه زیادی به دیتای مجموعه آموزشی نداشته باشد و مدل بیش از حد ساده شود و اصطلاحا Underfitting رخ دهد. همانطور که میدانیم یکی از علائم واضح underfit شدن مدل عملکرد ضعیف مدل در هر دو دیتای آموزشی و آزمون است. برای حل این مشکل میتوان در اولین قدم از یک مدل پیچیده تر استفاده کرد. یک راه حل دیگر استفاده از روش Cross Validation است برای مواقعی که میزان دیتا کم باعث میشود مدل نتواند یادگیری خوبی داشته باشد.

۲- پ)

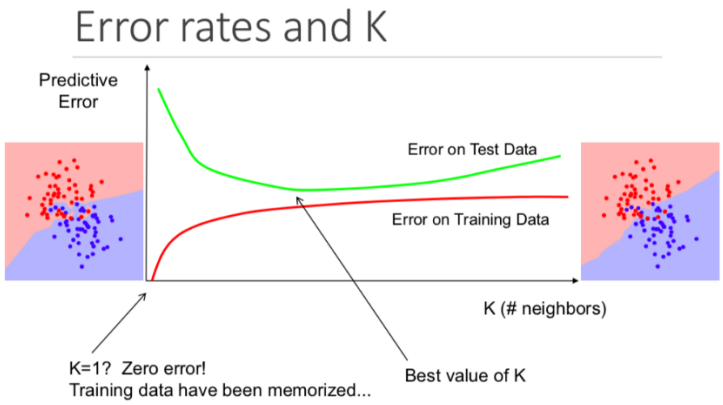
واریانس زیاد (High Variance) چطور قابل تشخیص است و برای مقابله با آن چه راه حلهایی وجود دارد؟

واریانس زیاد را در واقع میتوان نقطه مقابل بایاس زیاد قرار داد. هنگامی که واریانس زیاد باشد مدل سعی میکند به هر قیمتی بهترین عملکرد را روی مجموعه آموزشی داشته باشد. در این صورت مدل بر روی مجموعه آموزشی عملکرد بسیار خوبی دارد اما در عوض برای داده های تست عملکرد مدل بسیار ضعیف تر است که در این حالت اصطلاحا گفته میشود مدل overfit شده است. برای حل مشکل واریانس زیاد باید پیچیدگی مدل را با تکنیک های مختلفی کم کنیم. در اولین گام ممکن است بتوانیم خود مدل را ساده تر کنیم. اما روش هایی مثل منظم سازی و Dropout نیز میتوانند از پیچیدگی بیش از حد مدل جلوگیری کنند.

۲- ت)

یکی از الگوریتم هایی که در حوزه یادگیری ماشین مورد استفاده قرار میگیرد، الگوریتم نزدیک ترین همسایگی (KNN) است. برای مطالعه بیشتر درباره این الگوریتم میتوانید به این لینک مراجعه کنید. توضیح دهید که با تغییر مقدار K، بایاس و واریانس چه تغییری میکنند.

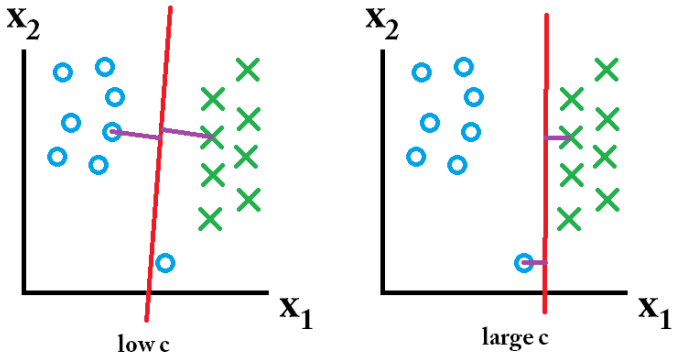
اگر مقدار پارامتر K را برابر ۱ قرار دهیم Bias صفر میشود ولی در عوض در مورد دیتای تست احتمال خطا بسیار زیاد میشود چون واریانس افزایش یافته است. در مقابل به مرور اگر مقدار K زیاد شود خطای آموزش (training error) زیاد میشود به دلیل افزایش Bias و خطای روی test set کاهش میابد (کاهش variance). میتوان اینگونه تعبیر کرد که با افزایش مقدار K تعداد همسایگان یک نقطه بیشتر میشود و عملا مدل پیچیده تری خواهیم داشت. شکل زیر تعابیر ذکر شده را با نمودارهای خطا روی مجموعه آموزشی و آزمون نشان میدهد.



۲- ث)

الگوریتم SVM یکی دیگر از الگوریتم های پرکاربرد حوزه یادگیری ماشین می باشد. توضیح دهید که با تغییر پارامتر C مقدار بایاس و واریانس چه تغییری میکنند. برای مطالعه بیشتر میتوانید به این لینک مراجعه کنید.

پارامتر C در الگوریتم SVM مشخص کننده میزان margin است. از طرفی هر چه margin بیشتر باشد میزان اطمینان به مدل بیشتر است و در طرف مقابل با افزایش margin میزان خطای مدل نیز میتواند افزایش یابد. پس در واقع انتخاب یک C مناسب توازن بین margin و خطا است. پس در نتیجه افزایش مقدار C باعث افزایش bias و کاهش C باعث افزایش variance میشود. شکل زیر یک مدل SVM را برای C های مختلف نشان میدهد که ارتباط C با bias و variance را بهتر نمایش میدهد



۲- ج)

منظم سازی پارامتر L1 و L2 را مقایسه کنید .

مهم ترین تفاوت بین این دو پارامتر منظم سازی این است که L1 تلاش میکند تا میانه داده ها را تخمین بزند در حالی که L2 با هدف جلوگیری از Overfit شدن مدل تلاش میکند تا میانگین داده ها را تخمین بزند. تفاوت دیگر این دو روش منظم سازی این است که L1 به انتخاب ویژگی (feature selection) کمک میکند و هنگامی که دیتاست ما دارای تعداد زیادی feature باشد تاثیر قابل توجهی در عملکرد مدل دارد.

۲- چ)

درست یا غلط بودن گزاره های زیر را مشخص کنید و دلیل پاسخ خود را نیز بیان کنید.

۲- چ -۱)

استفاده از منظم سازی، ممکن است باعث تضعیف عملکرد مدل شود .

درست – استفاده ی بیش از حد از منظم سازی باعث میشود مدل آموزش دیده،تبدیل به یک مدل خطی و یا حتی ثابت شود. به عبارتی دیگر با افزایش regularization توانایی مدل برای یادگیری training data کاهش میابد.

۲- چ – ۲)

اضافه کردن تعداد زیاد فیچرهای جدید، باعث جلوگیری از بیش برازش میشود .

غلط- تعداد بیش از حد feature ها باعث میشود مدل از روی مقدار یک فیچر خاص برای یک instance یک تصمیم گیری خاص کند که باعث میشود تعداد زیادی الگوی مشخص در مدل داشته باشیم و عملکرد روی training set بسیار خوب باشد و برای دیتای test نتایج بسیار ضعیفی بگیریم که به معنای overfit شدن مدل است.

۲- چ – ۳)

با زیاد کردن ضریب منظم سازی، احتمال بیش برازش بیشتر میشود.

غلط- همانطور که در پاسخ سوال ۲- چ – ۱ نیز گفته شد با افزایش ضریب منظم سازی مدل به سمت خطی شدن پیش میرود و ممکن است underfit شود. اگر ضریب منظم سازی کم شود ممکن است مدل overfit شود.

۲- ح)

فرض کنید یک مدل داریم که یکبار بدون منظم سازی و یکبار با منظم سازی آموزش داده میشود. کدام یک از دو مجموعه ضرایب زیر مربوط به منظم سازی و کدام یک بدون منظم سازی است؟ دلیل انتخاب خود را توضیح دهید .همچنین توضیح دهید که به نظر شما ضرایبی که با استفاده از منظم سازی به دست آمده است، مربوط به منظم -سازی پارامتر L1 است یا L2 ؟

- 13.3, 23.5, 53.2, 5.1
- 0.5, 1.2, 8.5, 0

به نظر میرسد اعداد سری دوم ضرایبی هستند که از منظم سازی استفاده کرده اند. همچنین با احتمال بیشتری از منظم سازی $L1$ استفاده شده است زیرا همانطور که در سوالات قسمت های قبل نیز ذکر شد پارامتر $L1$ تلاش میکند تا انتخاب ویژگی ها (حذف ویژگی هایی که کمتر موثر هستند یا اثر منفی دارند) را انجام دهد و در سری دوم اعداد مشاهده میشود یکی از مقادیر صفر شده است که ممکن است به دلیل استفاده از $L1$ باشد.

سوال ۳.

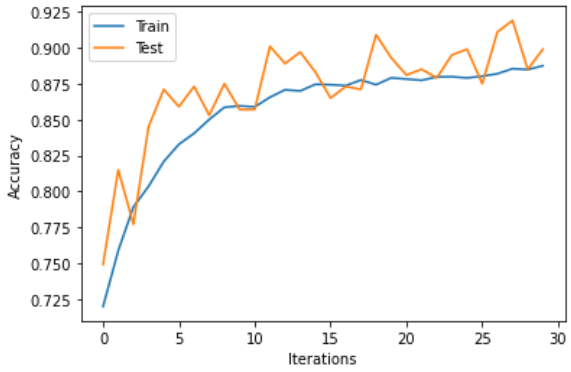
در تمرین قبل(تمرین سری سوم)سوال ۴ قسمت ب از شما خواسته شده بود که مدل را به نحوی تغییر دهید تا شبکه overfit شود. در این سوال قصد داریم با رو شهای بررسی شده در درس، مشکل به وجود آمده را برطرف کنیم.(برای این سوال باید از مدل overfit شده تمرین قبل که پیاده سازی کردهاید، استفاده کنید.)

سورس کد این سوال با نام HW4_Q3 ضمیمه فایل گزارش تمرین شده است. این سورس کد با توجه به قسمت های مختلف سوال ۴ قسمت مختلف دارد.

۱. **استفاده از data augmentation:**

در این قسمت از transform های کتابخانه Pytorch استفاده کردیم. بعد از بررسی حالت های مختلف داده افزایی، بهترین حالت برای این سوال و دیتاست حالتی بود که ۲ transformation کرنل گاوسی(با دو ابعاد ۵ و ۹ و سیگمای ۰.۱ و ۵) و affine transform های چرخش، جابه جایی و تغییر اندازه استفاده کنیم. پس از انجام این تغییرات برای اطمینان از یکسان بودن اندازه تمامی تصاویر، آن ها را با دستور resize به ابعاد ۲۸*۲۸ تبدیل کردیم.

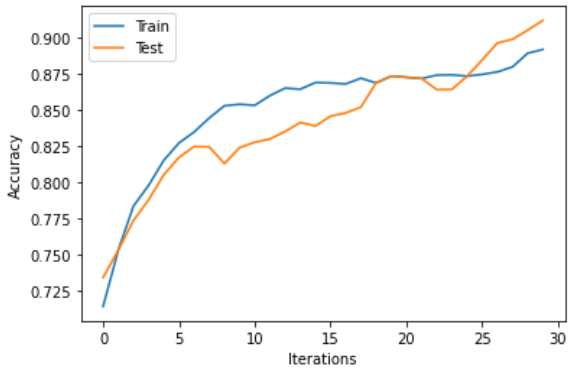
نمودار دقت برای training و test نشان میدهد این روش تاثیر بسیار مناسبی داشته است و کارکرد مدل افزایش یافته است و همینطور مدل overfit نشده است.



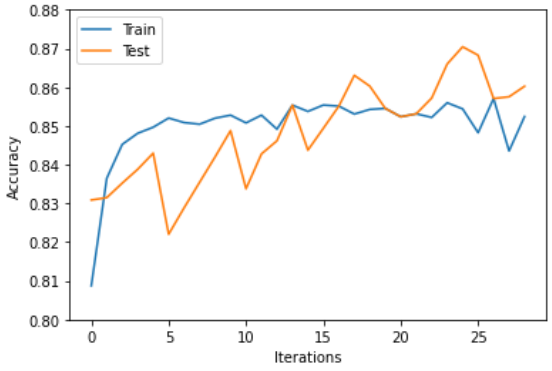
۲. **استفاده از منظم سازی L1 و L2:**

برای این قسمت ۳ نوع مدل پیاده سازی کردیم. حالت اول فقط L1 را اعمال کردیم. حالت دوم فقط L2 و در حالت سوم از هر دو پارامتر منظم سازی استفاده کردیم. نتایج نموداری حاصل از منظم سازی L1 به تنهایی به مراتب بهتر از حالتی است که از منظم سازی L2 و یا ترکیبی از هر دو پارامتر استفاده میکنیم است.

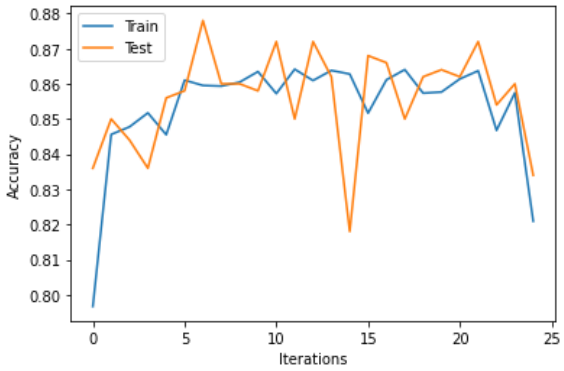
در حالتی که فقط از L1 استفاده میکنیم رفتار نمودار های accuracy تقریبا همانند استفاده از data augmentation است.



در حالتی که فقط از L2 استفاده میشود نیز بهبود کمی در مدل دیده میشود اما نه به قدرت L1. همچنین مدل بر روی داده ی test در طول مسیر رفتارهای متفاوتی از خود نشان میدهد(نوسان دارد).

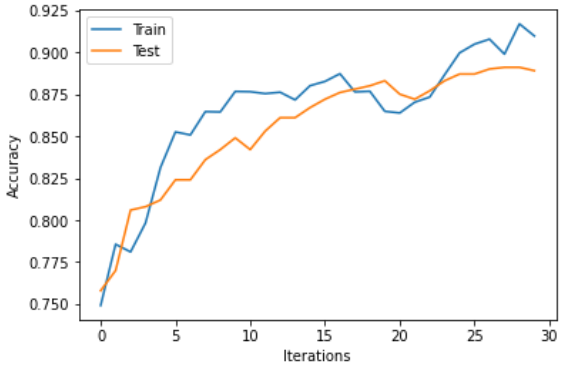


اما در حالتی که از هر دو منظم سازی L1 و L2 استفاده میکنیم بر خلاف انتظار مدل نتایج بسیار ضعیف تری ارائه میکند همچنین با گذشت زمان عملکرد مدل روی داده test کمی نسبت به ابتدا کاهش میابد.



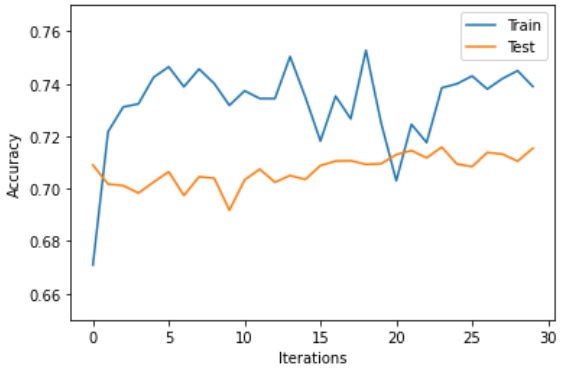
۳. **استفاده از Dropout:**

برای این سوال ما از dropout با احتمالات مختلف استفاده کردیم که پس از بررسی مشخص شد ۱۵٪ بهترین مقدار برای dropout برای این سوال است. نمودار دقت رفتاری مشابه L1 وdata augmentation دارد.



۴. **استفاده از ترکیبی از روش ها:**

برای این حالت، همه موارد استفاده شده در سه قسمت قبل را با هم ترکیب کردیم و نتایج خارج از انتظاری گرفتیم. در این حالت گمان میرفت که ترکیب روش های بالا منجر به مدلی بسیار قوی با عملکردی عالی شود اما نتایج حدود ٪۷۰ تا ٪۷۲ روی داده تست است و داده های train نیز با اختلافی ۲ الی ۳ درصدی بهتر از داده های test هستند.



در انتها جدول زیر برای جمع بندی و مقایسه حالات مختلف مدل آمده است.

Model	Training accuracy	Test accuracy
Data Augmentation	0.914	0.890
L1 regularization	0.892	**0.908**
L2 Regularization	0.860	0.865
L1 + L2 Regularization	0.804	0.834
Dropout (p = 0.15)	**0.915**	0.894
Combination	0.735	0.716

مراجع

1) https://towardsdatascience.com/understanding-the-bias-variance-tradeoff-165e6942b229

2) medium.com/hackernoon/tackle-bias-and-other-problems-solutions-in-machine-learning-models

3) medium.com/30-days-of-machine-learning/day-3-k-nearest-neighbors-and-bias-variance-tradeoff

4) stats.stackexchange.com/questions/203919/svm-does-c-increase-variance-or-stability-bias

5) https://medium.com/analytics-vidhya/l1-vs-l2-regularization-which-is-better-d01068e6658c

6) datascience.stackexchange.com/questions/76811/